Mitigating Statistical Bias within Differentially Private Synthetic Data

Sahra Ghalebikesabi[1], Harrison Wilde[2], Jack Jewson[2], Arnaud Doucet[1], and Sebastian Vollmer[4], Chris Holmes[1]

[1] University of Oxford, [2] University of Warwick, [3] Universitat Pompeu Fabra, [4] University of Kaiserslautern

UNIVERSITY OF OXFORD

ONE-MINUTE SUMMARY FOR BUSY RESEARCHERS

Inference on differentially private data is **biased**. This bias can be decreased by **differentially private importance sampling**.

1) DP synthetic data generation, i.e. with DP-GAN → 2) Importance weight estimation on synthetic data samples → 3) Downstream inference, i.e. classification, on weighted DP data

We propose three different approaches to estimate DP importance weights.

1. **Differentially private logistic regression:**
   (a) We learn a discriminating logistic regression to differentiate between true and synthesised data.
   (b) We privatise the coefficient vector by adding Laplacian noise.
   (c) We predict the importance weights by the sigmoid of the regression predictions on the synthetic data.
   (d) We correct the bias induced by the perturbations.
2. **Differentially private neural networks**: We train a discriminating neural network to differentiate between true and synthesised data using a modified DP-SGD procedure.
3. **Discriminator weights of DP-GANs**: If the synthetic data generator is a DP-GAN, we use the logit predictions of the discriminator on the synthesised data as importance weights.

More info!

1) Background

Differential privacy

- Two datasets $\mathcal{D}$ and $\mathcal{D}'$ are **neighbouring** when they differ by at most one observation. A randomised algorithm $g : \mathcal{M} \to \mathcal{R}$ satisfies (ϵ, δ)**-differential privacy** for $\epsilon, \delta \geq 0$ if and only if for all neighbouring datasets $\mathcal{D}, \mathcal{D}'$ and all subsets $S \subseteq \mathcal{R}$, we have
$$\Pr(g(\mathcal{D}) \in S) \leq \delta + e^{\epsilon} \Pr(g(\mathcal{D}') \in S).$$
- The **sensitivity** of g w.r.t a norm $|\cdot|$ is defined by the smallest number $S(g)$ such that for any two neighbouring datasets $\mathcal{D}$ and $\mathcal{D}'$ it holds that
$$|g(\mathcal{D}) - g(\mathcal{D}')| \leq S(g).$$

Dwork et al. (2006) show that to ensure the differential privacy of g, it suffices to add Laplacian noise with standard deviation $S(g)/\epsilon$ to g.

Importance weighting

Let the true data distribution be denoted by p_D, and the synthesised data be sampled from p_G. Additionally, we assume $w(x) := \frac{p_D(x)}{p_G(x)}$, and $p_G(\cdot) > 0$ whenever $h(\cdot)p_D(\cdot) > 0$. It then holds,
$$\mathbb{E}_{x \sim p_D}[h(x)] = \mathbb{E}_{x \sim p_G}[w(x)h(x)].$$
So we have almost surely the convergence
$$I_N(h|w) := \frac{1}{N_G}\sum_{i=1}^{N_G} w(x_i)h(x_i) \overset{N_G \to \infty}{\longrightarrow} \mathbb{E}_{x \sim p_D}[h(x)].$$
for $x_{1:N_G} \overset{\text{i.i.d.}}{\sim} p_G$. Note that we can use this approximation universally for *empirical risk minimisation*, and 2) *Bayesian updating*. If no weighting is possible, the importance weights can be used for *resampling* the synthetic data set before inference.

3) Selected Results

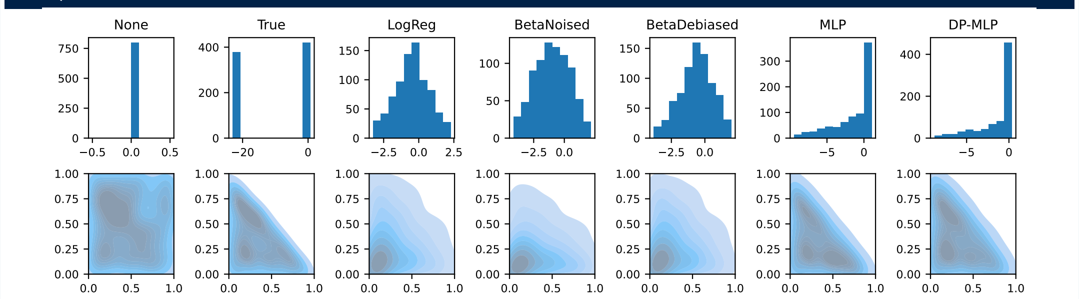

Figure 1. KDE plots of 100 observations sampled from a two dimensional uniform square distribution as SDGP (bottom left) and a uniform triangle distribution as DGP (second figure in second row). The first row depicts histograms of the computed weights starting with the true importance weights (True). The DP weights were privatised with $\epsilon = 1$, and the regularisation was chosen as $\lambda = 0.1$. The second row illustrates the importance weighted synthetic observations. We observe that while BetaDebiased corrects the weights of the logistic regression, the complex nature of the MLPs allows a better modelling of the DGP even in this simple setting.

IW	*β MSE* ↓	*MLP AUC* ↑
None	$0.6605_{\pm 0.03}$	$0.8502_{\pm 0.03}$
BetaNoised	$0.6247_{\pm 0.01}$	$0.8766_{\pm 0.00}$
BetaDebiased	$0.6240_{\pm 0.01}$	$\mathbf{0.8783_{\pm 0.00}}$
DP-MLP	$\mathbf{0.5813_{\pm 0.02}}$	$0.8683_{\pm 0.00}$
Discriminator	$0.6242_{\pm 0.01}$	$0.8631_{\pm 0.03}$
LogReg	$0.6234_{\pm 0.01}$	$0.8770_{\pm 0.00}$
MLP	$0.5707_{\pm 0.02}$	$0.8737_{\pm 0.00}$

Table 1. Mean and standard error over 10 runs with standard errors for $(\epsilon = 9.64, \delta =$

SDGP	*data*	*BetaNoised*	*BetaDebiased*
CGAN	Breast	$1.4833_{\pm 0.96}$	$\mathbf{0.0775_{\pm 0.01}}$
	Banknote	$0.0420_{\pm 0.02}$	$\mathbf{0.0413_{\pm 0.01}}$
	Iris	$8.7522_{\pm 4.98}$	$\mathbf{3.46_{\pm 1.30}}$
GAN	Housing	$8.2081_{\pm 7.77}$	$\mathbf{1.4406_{\pm 0.83}}$
DPCGAN	Breast	$0.0582_{\pm 0.01}$	$\mathbf{0.0445_{\pm 0.01}}$
	Banknote	$0.0420_{\pm 0.02}$	$\mathbf{0.0413_{\pm 0.01}}$
	Iris	$\mathbf{0.7834_{\pm 0.23}}$	$1.2300_{\pm 0.70}$

Table 2. Mean MSE of the DP log importance weights over 10 runs with standard errors reported in brackets

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

Zhanglong Ji and Charles Elkan. Differential privacy based on importance weighting. *Machine Learning*, 93(1):163–183, 2013.

2) Differentially Private Importance Weighting

Any calibrated classification method that learns to distinguish between data from the true data distribution, labelled thenceforth with $y = 1$, and from the synthetic data distribution, labelled with $y = 0$, can be used to estimate the likelihood ratio (Sugiyama et al., 2012). We compute
$$\widehat{w}(x) = \frac{\widehat{p}_D(x)}{\widehat{p}_G(x)} = \frac{\widehat{p}(x|y=1)}{\widehat{p}(x|y=0)} = \frac{\widehat{p}(y=1|x)}{\widehat{p}(y=0|x)}\frac{\widehat{p}(y=0)}{\widehat{p}(y=1)}$$
where $\widehat{p}$ are the probabilities estimated by such a classification algorithm.

Differentially private logistic regression

- If the data is scaled to a range from 0 to 1 such that $X \subset [0,1]^d$, Chaudhuri et al. (2011) show that the L_2 sensitivity of the optimal coefficient vector estimated by $\widehat{\beta}$ in a regularised logistic regression with model
$$\widehat{p}(y=1|x_i) = \sigma(\widehat{\beta}^T x_i) = \left(1 + e^{-\widehat{\beta}^T x_i}\right)^{-1}$$
is $S(\widehat{\beta}) = 2\sqrt{d}/(N_D\lambda)$ where λ is the coefficient of the L_2 regularisation term added to the loss during training.
- Ji and Elkan (2013) propose to compute DP importance weights by training such an L_2 regularised logistic classifier on the private and the synthetic data, and perturb the coefficient vector $\widehat{\beta}$ with Laplacian noise. For a d dimensional noise vector ζ with $\zeta_j \overset{i.i.d.}{\sim}$ Laplace$(0, \rho)$ with $\rho = 2\sqrt{d}/(N_D\lambda\epsilon)$ for $j \in \{1, \ldots, d\}$, the private regression coefficient is then $\overline{\beta} = \widehat{\beta} + \zeta$, and
$$\log \overline{w}(x_i) = \overline{\beta}^T x_i = \widehat{\beta}^T x_i + \zeta x_i. \tag{1}$$

Proposition 1 *(informal)* Let $\overline{w}$ denote the importance weights computed by noise perturbing regression coefficients as in Equation 1 (Ji and Elkan, 2013, Algorithm 1). The resulting IS estimator is biased.

Proposition 2. Let $\overline{w}$ denote the importance weights computed by noise perturbing the regression coefficients (Ji and Elkan, 2013, Algorithm 1). Define
$$b(x_i) := 1/\mathbb{E}_{p_\zeta}[\exp\left(\zeta^T x_i\right)],$$
and adjusted importance weight
$$\overline{w}^*(x_i) = \overline{w}(x_i)b(x_i) = \widehat{w}(x_i)\exp\left(\zeta^T x_i\right)b(x_i).$$
The importance sampling estimator $I_N(h|\overline{w}^*)$ is unbiased and (ϵ, δ)-DP for $\mathbb{E}_{p_\zeta}[\exp\left(\zeta^T x_i\right)] > 0$.

Differentially private neural networks

We train a discriminating neural network, with following SGD variant.

Input: Examples $x_{1:N_D}, y_{1:N_D}$ from the DGP and $x_{N_D+1:N_D+N_G}, y_{N_D+1:N_D+N_G}$ from the SDGP, loss function $\mathcal{L}(\theta) = \frac{1}{N_G+N_D}\sum_i \mathcal{L}(\theta, x_i, y_i)$. Parameters: learning rate η_t, noise scale σ, expected lot size L, gradient norm bound C.

Initialise θ_0 randomly
for $t \in [T]$ **do**
  Construct a random subset $L_t \subset \{1, \ldots, N_D + N_G\}$ by including each index independently at random with probability $\frac{L}{N_D+N_G}$
  Compute gradient For each $i \in L_t$, compute $g_t(x_i, y_i) \leftarrow \Delta_{\theta_t}\mathcal{L}(\theta_t, x_i, y_i)$
  Clip gradient $\overline{g}_t(x_i, y_i) \leftarrow g_t(x_i, y_i)/\max(1, \frac{||g_t(x_i,y_i)||_2}{C})$
  Add noise $\tilde{g}_t \leftarrow \frac{1}{L}\sum_{i \in L_t}(\overline{g}_t(x_i, y_i) + N(0, \sigma^2 C^2 \mathbf{I})\mathbb{1}_{(y_i=1)})$, where $\mathbb{1}_{(y_i=1)}$ is 1 if $y_i = 1$ and 0 otherwise
  Descent $\theta_{t+1} \leftarrow \theta_t + \eta_t \tilde{g}_t$

Output: θ_T and the overall privacy cost (ϵ, δ) using the moment's accountant of Abadi et al. (2016) with sampling probability $q = \frac{L}{N_D+N_G}$.

Algorithm 1. Relaxed DP SGD; differences to Abadi et al. (2016) in blue.

Discriminator weights of DP-GANs

- GANs produce realistic synthetic data by trading off the learning of a generator Ge to produce synthetic observations, with that of a classifier Di learning to correctly classify the training and generated data as real or fake.
- In contrast to the weights computed from DP classification networks, this approach is more robust, requires less hyperparameter tuning, and does not use up additional privacy budget!